DOCUMENT RESUME

ED 325 479 TM 015 630

AUTHOR Baghi, Heibatollah; Ferrara, Steven F.
TITLE Detecting Differential Item Functioning Using IRT and Mantel-Haenszel Techniques: Implementing Procedures and Comparing Results.
PUB DATE Feb 90
NOTE 35p.; Paper presented at the Annual Meeting of the Eastern Educational Research Association (Clearwater, FL, February 14-17, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Chi Square; Comparative Analysis; Correlation; Grade 9; *High School Freshmen; High Schools; *Item Bias; Item Response Theory; Testing Problems; *Test Items; Test Reliability
IDENTIFIERS *Mantel Haenszel Procedure; *Maryland Test of Citizenship Skills; Three Parameter Model

ABSTRACT

Two techniques for detecting differential item functioning (DIF) of test items are compared: (1) item response theory (IRT), using the three-parameter model; and (2) Mantel-Haenszel chi square techniques (MHCS). The steps necessary for identifying differentially functioning items are identified for both approaches. Using data from the 1988 Maryland Test of Citizenship Skills (MTCS), the following parameters were investigated: (1) the stability of the MHCS statistic across sample sizes; (2) the stability of the DIF index (MH Alpha) in the MHCS approach across several score groups; (3) the correlation between IRT DIF indices and MH Alpha; and (4) agreement between the IRT and MHCS techniques in identifying biased items. The MTCS was administered to about 50,000 ninth-graders in January and February of 1988. For IRT purposes, random comparison groups of 1,000 students each were created, while the MHCS procedure used samples of 1,000, 750, 500, and 200 examinees. The IRT DIF procedure identified four items in male/female comparisons and three items in white/black comparisons in the MTCS that have significant unequal probabilities of a correct response. While correlation results indicate that the two techniques disagree regarding item bias, agreement in terms of total hits indicates that the MHCS technique is an adequate substitute for the three-parameter IRT approach if the sample size is at least 750. In terms of the percentage of items identified as biased, the techniques appear to be equally good in detecting DIF. Five graphs and 11 tables present study data. A 35-item list of references is provided. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED325479

Detecting Differential Item Functioning Using IRT and Mantel-Haenszel Techniques: Implementing Procedures and Comparing Results

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

HEIBATOLLAH BAGHI
STEVEN F. FERRARA

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
HEIBATOLLAH BAGHI
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

DIVISION OF INSTRUCTION

Maryland State Department of Education
200 West Baltimore Street
Baltimore, Maryland 21201

Paper presented at the
Annual Conference of the
Eastern Educational Research Association

Clearwater, Florida
February 1990

TM015630

The study of test items that function differently for subpopulations of examinees is of concern to test developers as well as test information consumers. This concern is especially critical in high stakes competency-based testing, where graduation certification is contingent on passing one or more tests. The phenomenon of items that function differently for different examinee subpopulations -- that is, that may be easier or more difficult for one group than another -- is referred to as differential item functioning (DIF). DIF was originally called item bias. Researchers have attempted to define and detect DIF in various ways. In educational measurement, the term "bias" is used in reference to tests and their use, usually for selection and classification, or individual items and their effect on total test scores. Test items may be considered biased when a focal group scores disproportionately lower or higher than a reference group due to factors other than ability differences.

Clearly, the definition of item bias is dependent in part upon the techniques that are used to find differentially functioning items. For example, Angoff and Ford (1973) considered an item biased if the item difficulty index is significantly higher or lower in one group than in another group. Scheuneman (1979) considered an item biased if, for all examinees having the same score on a test that includes that item, the proportion of examinees answering the item correctly is substantially different for various subpopulations being considered. When using item response theory (IRT) to investigate item bias, an item is considered biased if the item characteristic curves (ICCs) for an item are not identical for compared subpopulations (see Crocker & Algina, 1986, p. 377). In IRT, individuals from different subpopulations with the same ability should have equal probabilities of responding correctly to unbiased items. Item bias is defined in chi-square techniques somewhat similarly to the IRT definition of item bias. In chi-square techniques an item is considered biased if, within groups of individuals with total test scores in the same test score interval, the proportion of individuals responding correctly to the item is different for different subpopulations (Crocker & Algina, 1986, p. 383). A widely accepted definition for DIF is that an item is considered biased if examinees with equal ability, but from different subpopulations, have unequal probability of responding correctly to an item (Angoff, 1982).

Many studies of DIF techniques in the past several years have been devoted to comparing different techniques. The numerous techniques proposed for detecting DIF have been

narrowed down in recent years to several of the most promising. There exist several comprehensive reviews of the DIF literature (Burrill, 1982; Ironson, 1982; Rudner, Getson, & Knight, 1980; Osterlind, 1987; and Shepard, Camilli, & Williams, 1985). The consensus from this research is that "the ICC approach is the most generally valid of all biased item detection methods" (Osterlind, 1987, p. 69). IRT techniques are the theoretically preferred procedures for detecting DIF because they least confound real mean differences in group performance with bias (Lord, 1977). The sample invariance property of IRT provides a theoretical framework of how DIF is defined and detected in a test. ICCs describe the relationship between the probability of responding correctly to an item and examinee ability (i.e., theta) measured by the test. To study DIF, ICCs are estimated separately for each group. If an item functions similarly in two comparison groups, then the probability of a correct response should be the same for examinees of equal ability.

Although the IRT approach is theoretically the most sound approach to detecting DIF, according to Shepard and her associates (1981), it has accompanying difficulties. IRT analyses require large amounts of computer time to obtain satisfactory solutions and can be difficult for practitioners to apply. In addition, the three-parameter model requires a minimum of 1,000 cases per group (i.e., for LOGIST) to estimate item parameters, a requirement that often is difficult to meet in minority samples. As a result, other techniques have been developed; for example, the Mantel-Haenszel chi-square (MH chi-square) technique that is considered to produce approximations to item response theory DIF results (Scheuneman, 1979; Holland & Thayer, 1986). An advantage of the MH chi-square technique is that it is easier to apply than the IRT technique and apparently does not require large sample sizes. In contrast to Shepard's and others' view on the superiority of IRT DIF approaches, other consider the MH chi-square technique the most powerful and unbiased test against the alternative hypothesis that probabilities of correct responses are uniform across ability levels (Holland & Thayer, 1988). However, the relationship between the MH chi-square indices and (a) sample size, and (b) the number of score groups has not received adequate attention. More empirical research is needed concerning this aspect of the MH chi-square technique and its agreement with IRT approaches in detecting specific differentially functioning items.

In this paper we describe briefly and compare two techniques for detecting DIF: IRT (using the three-parameter model), and MH chi-square techniques. First, we describe both procedures, including the steps necessary to identify differentially functioning items using the IRT three-parameter approach and the MH chi-square approach. Last, we

determine the (a) stability of the MH chi-square statistic across sample sizes, (b) stability of MH Alpha (the DIF index in the MH chi-square approach) across number of score groups, (c) correlation between IRT DIF indices and MH Alpha, and (d) agreement between the IRT and MH chi-square techniques in identifying biased items.

IRT and MH Chi-square DIF Techniques, Methods of Analysis, and Procedures

Three-Parameter IRT DIF Technique and Procedures

Using IRT methods, an item is judged unbiased if examinees with the same ability level, but of different group membership, have equal probabilities of a correct response. In other words, an item is unbiased if the ICCs are identical for both subpopulations. ICCs like those shown in Figure 1 describe the relationship between the probability of a correct response $P(U_g=1/O)$, and increasing amounts of the ability (theta) measured by a test.

Figure 1 about here

Total observed scores are used as estimates of theta and the probability of responding correctly to item g given an examinee of ability theta is estimated using Birnbaum's three-parameter model (Birnbaum, 1968):

$$P(ug=1/theta_i) = c_g + (1-c_g)\ [1-exp(-1.7(theta-b))]^{-1}, \quad (1)$$

where a_g is item discrimination, b_g is item difficulty, and c_g is an estimate of guessing. ICCs are defined by these three parameters.

ICCs for a biased item are illustrated in Figure 2. These curves are constructed separately for each comparison group using the above formula and based on responses by the different groups to an item. These curves are not identical because the location parameters for the two ICCs are not equal. This item is judged biased because examinees of the same ability level but from different groups do not have similar proportions of correct responses. This method simultaneously analyzes differences in item discrimination, difficulty, and guessing.

Figures 2 and 3 about here

As shown in Figure 3a-3c, bias may be solely in one parameter or in a combination of different parameters. For

example, in Figure 3a the curves cross and create two areas of difference because the discrimination (a_g) parameters are not equal. In Figure 3b, the difficulty (b_g) parameters are not equal, this creating a single area of difference between the two curves. The ICCs in Figure 3c are not parallel because both the a_g and b_g parameters are not equal. There are many IRT bias indices that are intended to quantify these differences in ICCs (see, for example, Hills, 1989). One method proposed by Rudner (1977), and used in this study, is to compute areas of difference between two ICCs. The size of this area will be low for relatively unbiased items and high for relatively biased items. The area can be approximated by using the formula:

$$A_g = \sum_{-3}^{3} [P_{1g}(\text{theta}) - P_{2g}(\text{theta})] \Delta \text{ theta}, \qquad (2)$$

$$\text{where } \Delta \text{theta} = 0.01$$

In this equation, P_{1g}(theta) and P_{2g}(theta) refer to values of ICCs for each of two comparison groups. Values of P_{1g} are calculated separately for each group and for each value of theta from -3.00 to 3.00 (i.e., 3 standard deviations above and below the mean) in increments of 0.01. Because of the absolute value sign, negative values of P_{1g}(theta) - P_{2g}(theta) are converted to positive values. This conversion is desirable when the ICCs for the two subgroups cross each other; without this conversion the positive and negative differences would balance out to some extent, even if items are biased.

Before the DIF index is calculated, item parameter estimates for each group must be converted to the same scale. This re-scaling is necessary when item and person parameter estimation procedures standardize on theta, the ability score, rather than on b_g. In this procedure, the estimate of theta is set separately for each group. In each group, the theta score estimates have a mean of zero and a standard deviation of 1.0. Because the theta scale estimates for two groups are different, the scales for the estimates of item difficulty and discrimination are different and must be equated. In this study we linearly transformed the difficulty and discrimination parameters separately (see Crocker & Algina, 1986, p. 379).

Differences between ICCs. Examination of differences between three-parameter ICCs involves six steps which we describe below.

1. Estimate item parameters for all items for two random groups using the PC-BILOG program (Mislevy & Bock, 1986) or some other appropriate item calibration program. DIF results from analysis of these parameters provide a criterion for distinguishing real DIF, which may be caused

by some form of bias against a subgroup, and apparent DIF dur to sampling error.

2. Estimate item parameters for all items separately for each reference and focus group.

3. Linearly transform item parameters for the reference and focus groups to the same scale using the following steps:

.The mean and the standard deviation of the item difficulties of items for each group were calculated (M_F, SD_F, M_R, SD_R)

.Linear parameters A and B were calculated as follows:

$$A = (SD_R/SD_F)$$

$$B = M_R - (A * M_F)$$

.Item parameter estimates were adjusted using the following equations:

$$b_R = A * b_f + B$$
$$a_R = a_f\ (1.00/A)$$

where a = item discrimination index and b = item difficulty value.

4. Using the item parameter estimates for the two random groups as input, calculate differences between the two ICCs for each item.

5. Calculate the absolute difference between the ICCs for each item and the mean, standard deviation, and 99 percent confidence interval for these absolute differences.

6. Use the 99% confidence intervals from step 4 as a baseline to identify extreme differences between ICCs for reference and focus groups; any difference not contained in the confidence interval is considered an indication of DIF.

MH Technique and Procedures

The MH technique for DIF analysis , or MH chi-square technique, described by Holland and Thayer (1988), has also been applied to the item data in this study. The analysis is accomplished in six steps, which we describe below.

1. Create K 2X2 tables, where K is some number of score categories.

2. Create a 2X2 contingency table, as illustrated in Table 1.

Table 1 about here

---

3. Use the K 2X2 tables for any one item to calculate the Mantel-Haenszel chi-square statistic:

$$\text{MH CHISQ} = \frac{(|\Sigma A_j - E(A_j)| - 0.05)}{\Sigma \text{Var}(A_j)} ,$$

where A_j = number of examinees in the reference group that answered correctly,

B_j = the number who answered incorrectly,

Cj and Dj are defined similarly, but for the focus group,

$nR_j = A_j + B_j$,

m_{1j} and m_{0j} are the numbers of examinees in score group j that passed and failed an item,

$E(A_j) = (nR_j)\ (m_{1j})/T_j$, and

$$\text{Var}(A_j) = \frac{nR_j\ nF_j\ m_{1j}\ m_{0j}}{(T_j)\ (T_j - 1)} .$$

The MH CHISQ statistic indicates statistically significant differences between the comparison groups.

4. Calculate MH Alpha, the common odds-ratio across the K 2X2 tables for each item:

$$\text{MH Alpha} = \frac{\Sigma\ A_j D_j / T_j}{\Sigma\ B_j\ C_j\ /\ T_j}$$

MH Alpha indicates "practical significance". An MH Alpha value of 1.0 indicates no DIF. If the MH Alpha is greater than 1.0, the item appears to be functioning differentially in favor of a reference group; if it is less than 1.0 then it is functioning differentially in favor of a focal group.

5. Calculate MH Delta:

$$\text{MH Delta} = -(2.35)\ \ln\ (\text{MH Alpha})$$

MH Delta transforms the MH Alpha statistic to obtain a measure of DIF expressed in terms of differences on the

Educational Testing Service (ETS) delta scale (Holland & Thayer, 1985). MH Delta indicates the average amount by which examinees in a focal group found an item more difficult than did a reference group. Negative values of MH Delta indicate items that favor a reference group; positive values indicate items that favor a focal group.

6. Compare MH CHISQ to the chi-square distribution with one degree of freedom to test the null hypothesis of no differential item functioning. This statistic is considered to be the most powerful unbiased test of DIF (Cox, 1970).

In the final sections of this paper we describe results from implementing the IRT and MH chi-square techniques in the Maryland Test of Citizenship Skills (MTCS) and draw conclusions about the consistency of the results within and across techniques.

Method and Procedures for This Study

Data Source

The data for this study are item responses on one form from the 1988 edition of the Maryland Test of Citizenship Skills (MTCS). The MTCS consists of 45 multiple-choice items that assess students' knowledge and skills in three domains: Constitutional Government; Principles, Rights, and Responsibilities; and Politics and Political Behavior. Annual forms of the MTCS are constructed by sampling items from a large bank that has been calibrated using the Rasch model. Students must pass the MTCS, along with three other minimum-competency tests, in order to receive a Maryland high school diploma. The Maryland Functional Testing Program (MFTP) uses two approaches for detecting DIF: judgmental reviews and statistical analysis. Before newly written items are field tested, experts in ethnic and sex bias review item language and the situations they pose for potential sources of bias. Flagged items are examined for potential causes of DIF before they are included in a student's score on the MTCS, and later resubmitted for review by bias specialists.

Construction of Student Samples

The MTCS was administered to approximately 50,000 9th grade students during January and February, 1988. For IRT analyses, random comparison groups (referred to as "random 1" and "random 2") of 1,000 students each were created by randomly selecting cases from the entire pool. Male-female and white-black comparison groups were created by randomly sampling cases from within sex strata. Males and whites are treated in this analysis as the "reference" group; females blacks are treated as the "focal" group. (Any theoretically

or practically relevant groups can be paired as reference and focal groups in a DIF analysis.) Analyses for the MH chi-square approach is performed in randomly representative samples of 1,000, 750, 500, and 200 male, female, white, and black students.

Test Unidimensionality

It is assumed in DIF analyses that the test under scrutiny is unidimensional; that is, that all items measure the same latent ability, skills, and knowledge (see Hills, 1989). Investigating the unidimensionality assumption is problematic because researchers often do not agree on appropriate methodology and criteria for testing this assumption. In this study, the recommendations of Reckase (1979) for determining unidimensionality of a test were used as follows: (a) In a factor analysis of test items, the first unrotated principal component should account for at least 20 percent of total test variance; and (b) The eigenvalue for the first principal component should be large relative to the eigenvalue for the second principal component.

Procedures

The IRT procedure was implemented only in the samples of 1,000 examinees, as described above. The MH chi-square procedure was implemented in the samples of 200, 500, 750, and 1,000 examinees, and using 4, 5, 6, 7, 8, and 10 score groups, to investigate the effects of sample size and numbers of score groups on results. The two DIF techniques were compared in terms of correlation between IRT and MH Alpha DIF indices and of agreement in detecting specific differentially functioning items.

Unlike other DIF studies that use simulated data, there is no means in this study for pre-identifying biased items. Instead, we identify biased items using the three-parameter IRT technique and compare items identified in the MH method. Previous research (Shepard et al., 1985) indicates that using the three-parameter IRT technique produces superior results in both real and simulated test data.

Results

Unidimensionality

The unidimensionality of the MTCS was tested by extracting principal components from item correlation matrices computed separately on the random1 and random2 samples of male, female, white, and black students. The

component for the male group was 20 percent, for the female group 19 percent. The eigenvalue of the first principal components for each of these groups was at least four times as large as the next largest component. Similar results were found for white and black groups. Further, scree plots of the eigenvalues in Figures 4 and 5 for the male-female and white-black groups suggest a unidimensional solution. It may be concluded from these results that the items in the 1988 MTCS are reasonably unidimensional for all four subgroups.

Figures 4 and 5 about here

Three-Parameter IRT DIF Analysis

Tables 2 and 3 present estimates of item discrimination, difficulty, and guessing for comparison groups: the two random groups (random1 and random2, which appear in both tables), the male and female groups, and the white and black groups.

Tables 2 and 3 about here

Using these item parameter estimates as input, differences between ICCs for each item were calculated for each compared subgroup. Differences between ICCs for pairs of groups were examined by comparing the area between the ICCs for the two independent random samples to the area between the ICCs for the male-female and white-black samples. This method takes into account differences between compared groups in item parameter estimates simultaneously. The confidence interval for the absolute differences between the two random groups was used as a baseline to identify extreme differences in the ICCs found in the male-female and white-black samples. Any differences not contained in these confidence intervals are considered an indication of a differentially functioning item. ICC differences, means, standard deviations, and 99 percent confidence intervals are reported in Tables 4 and 5 for the male-female and black-white groups. As Table 4 shows, only four out of 45 items in the male-female comparisons are outside of the confidence interval. Table 5 shows that only three out of 45 items in the black-white comparisons are outside of the confidence interval. These items are identified as differentially functioning items and were flagged for additional review by the bias review committee.

Tables 4 and 5 about here

MH Chi-square DIF Analyses

Stability of MH CHISQ across sample sizes. Tables 6 and 7 show MH CHISQ, MH Alpha, and MH Delta statistics for male-female and white-black comparisons. The MH CHISQ statistic is compared to the chi-square distribution with one degree of freedom. MH Alpha and MH Delta indicate the direction of any bias. The results in Tables 6 and 7 indicate a pattern in the MH CHISQ statistic: as sample size decreases fewer items are identified as biased. These results indicate a dependence between the size of the chi-square statistic and the sample size used in analysis. For example, in Table 6 in the sample size of 200 no item was identified as biased. However, in the sample size of 1,000 seven items were identified as biased. Four out of those seven items were the items identified in the IRT DIF procedure (see Table 4). It should be noted that the number of score groups (k) used was seven in the analyses for Tables 6 and 7 because there were few examinees at the lower end of the score distribution. (Originally, 10 score groups were used, lower score intervals were collapsed, resulting in seven score groups). This collapsing was selected because it minimized the number of unmatched focal group members.

Tables 6 and 7 about here

Stability of MH Alpha across numbers of score groups. Tables 8 and 9 show the Spearman rank correlations of MH Alpha statistics obtained from different numbers of score groups, within sample sizes, for male-female and white-black groups. All correlations are above .90. These results suggest that using as few as four groups and sample sizes as small as 200 may be "optimal" for estimating MH Alpha.

Correlations between IRT and MH Chi-square DIF indices. Tables 8 and 9 also contain correlations between IRT and MH Alpha DIF indices across sample sizes and numbers of score groups and for male-female and white-black subgroups. In general, correlations between MH Alpha and the IRT DIF index (absolute differences between ICCs) are low. The highest correlation, 0.514, is the correlation for the sample size of 1,000 for the male-female comparison (see Table 8). This result agrees with previous research by Skaggs and Lissitz (1988) that indicates that results from two different techniques (IRT and MH) of DIF analysis often do not agree very well.

Tables 8 and 9 about here

Comparison of IRT and MH Chi-Square DIF Analysis Results

Percentage of agreement between IRT and MH chi-square results. Correlations are only a crude measure of how well MH chi-square results agree with IRT DIF results. Also important is the accuracy with which the two DIF techniques identify biased and unbiased items, using the IRT DIF technique as the criterion. Percentages of agreement on items identified by the IRT DIF index and the MH chi-square techniques, using different sample sizes and for male-female and white-black subgroups, appear in Tables 10 and 11. Percentages of items identified as both "biased" and "unbiased" (i.e., total hits) and proportions identified as "biased" (true positives) are reported.

Tables 10 and 11 about here

Total hit agreement rates range from 84 to 91 percent. An interesting pattern is apparent in true positive rates. As sample size increases, percentages of true positives increase and correctly identify more items as biased. For example, out of four items identified as biased in the IRT procedure, the same three items are identified as biased by MH chi-square technique in the sample size of 1,000.

Conclusion

The IRT DIF procedure used in this study identified four items in male-female comparisons and 3 items in white-black comparisons in the Maryland Test of Citizenship skills which have significant unequal probabilities of a correct response (i.e., ICCs). Further inspection of item parameter estimates shows that three out of four items were differentially more difficult for the focal groups analyzed here. Identifying only a small number of items is not surprising considering the extensive judgmental and empirical reviews that items undergo prior to operational use.

The MH chi-square technique was implemented in representative samples of different sizes. Results indicate that as sample size decreases the MH CHISQ statistic identifies fewer items as biased. For example, in the sample size of 200, no item was identified as biased while seven items were identified in the sample of 1,000 examinees. However, four out of seven items were identified as biased using IRT procedure. MH Alpha correlations are low in different sample sizes and with the IRT DIF index. In contrast, MH Alpha correlations are high in different numbers of score groups within sample size. These results

numbers of score groups within sample size. These results suggest that having as few as four groups and samples of 200 may be adequate for estimating MH Alpha. This conclusion supports previous research by Raju, Bode, and Larsen (1989).

Correlations between the IRT DIF index (absolute differences betwenn ICCs) and MH Alpha indicate that the two techniques do not agree regarding item bias. This conclusion supports previously published research (Skaggs & Lissitz, 1988). In this study, a Marginal Maximum Likelihood (MML) procedure was used to estimate IRT item parameters, while in the previous research a Joint Maximum Likelihood (JML) procedure was employed. MML estimators are consistent estimators (Baghi, 1988), an advantage over less consistent JML estimates (see Lord, 1986) that may impact DIF analysis results.

Agreement between MH chi-square and IRT DIF indices (in terms of total hits) indicates that the MH chi-square technique is an adequate substitute for the three-parameter IRT approach, if sample size is at least 750. In addition, in terms of percentages of items identified as biased, both the MH chi-square technique in samples of 1000 appear equally as good as the IRT approach in detecting differentially functioning items.

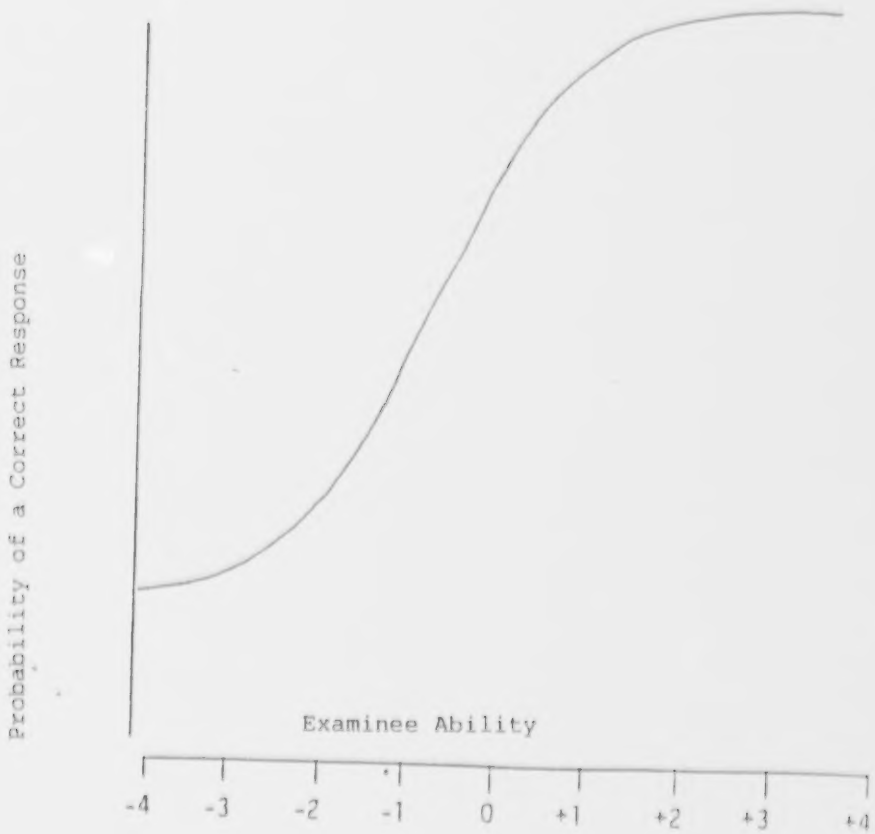

Figure 1 A hypothetical ICC

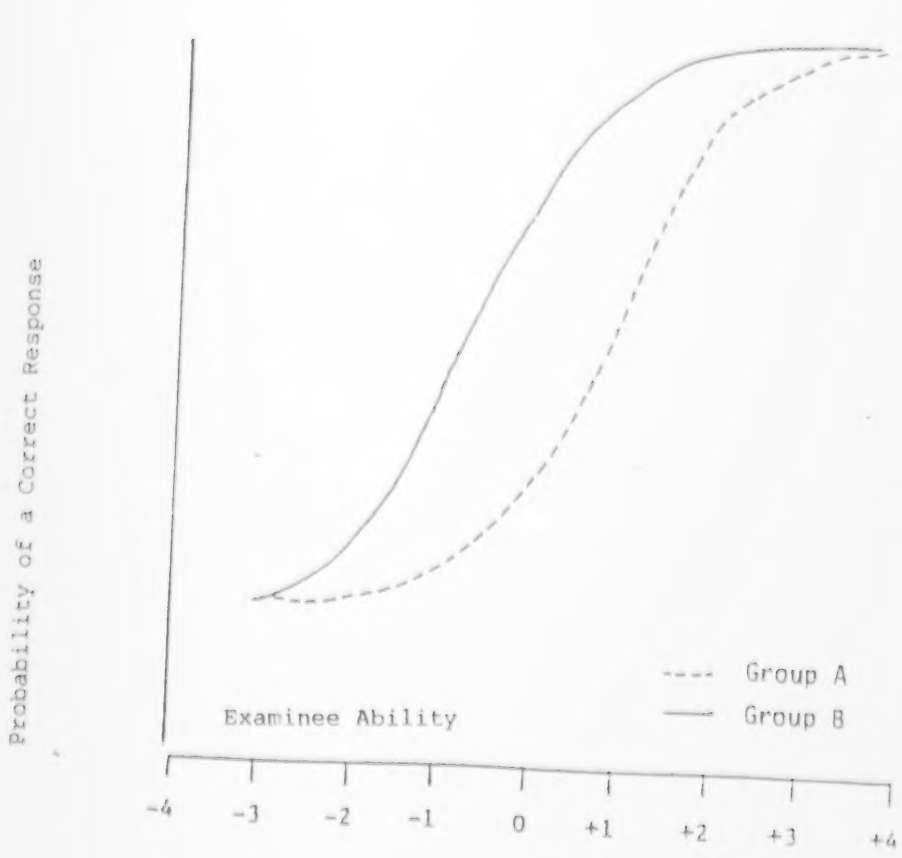

Figure 2. Two hypothetical response distributions

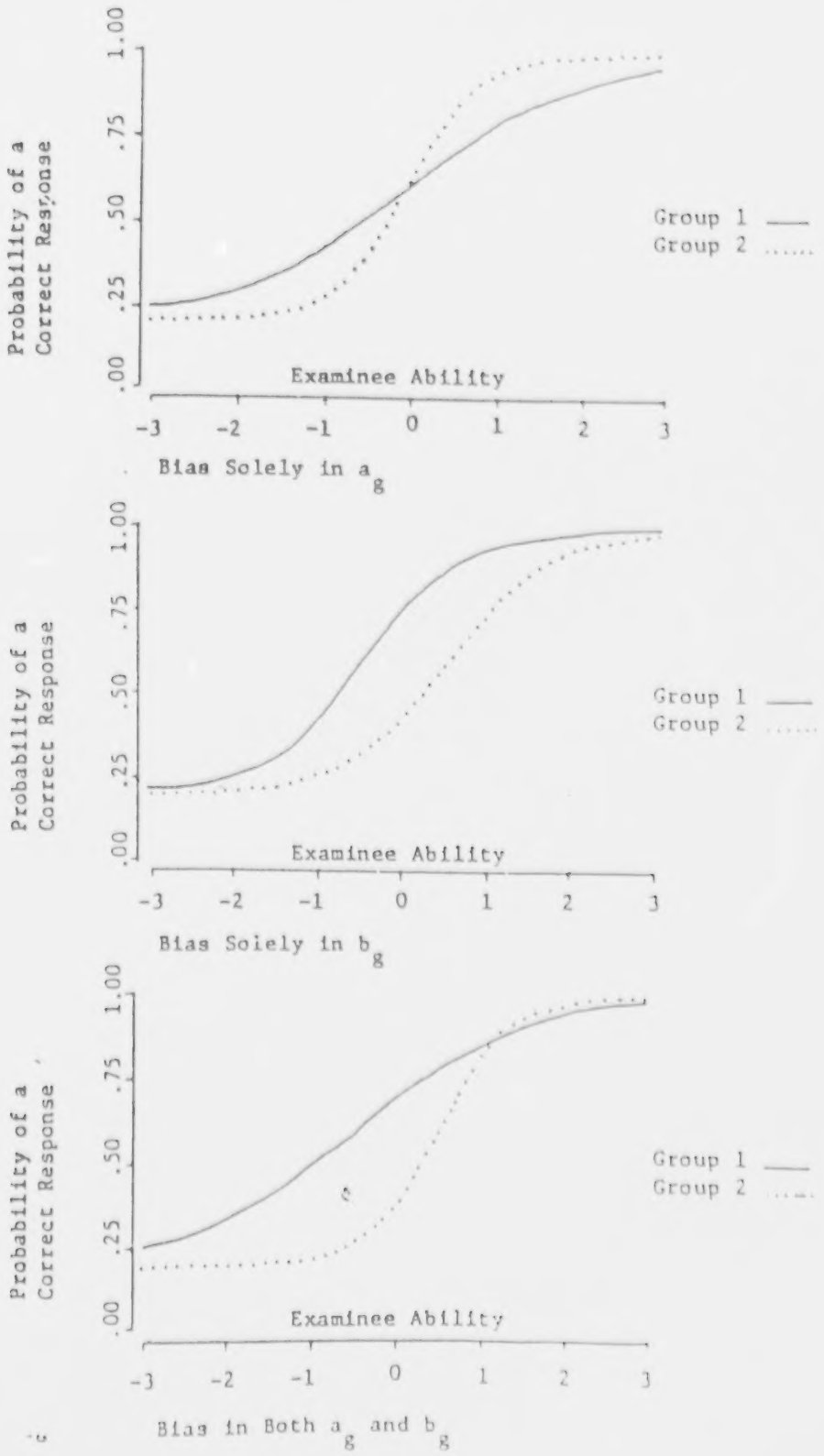

Figure 3 Bias in hypothetical ICCs

From: (Merz & Neal 1978)

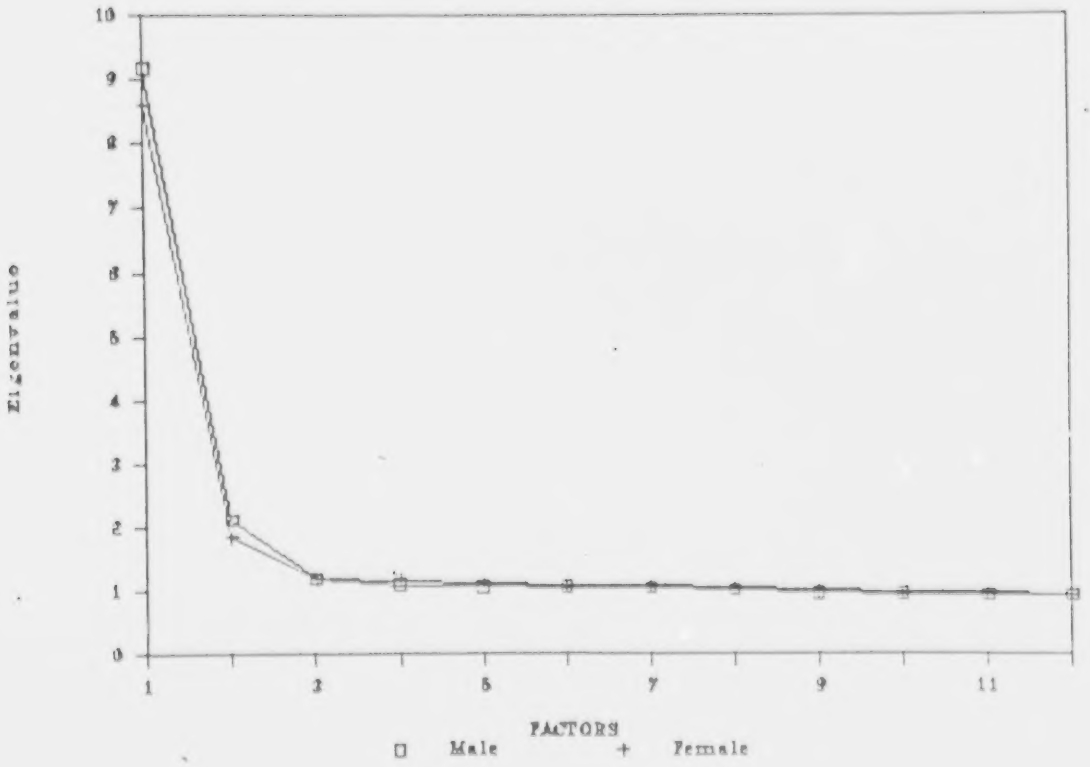

Figure 4. Scree Plot of Eigenvalues.

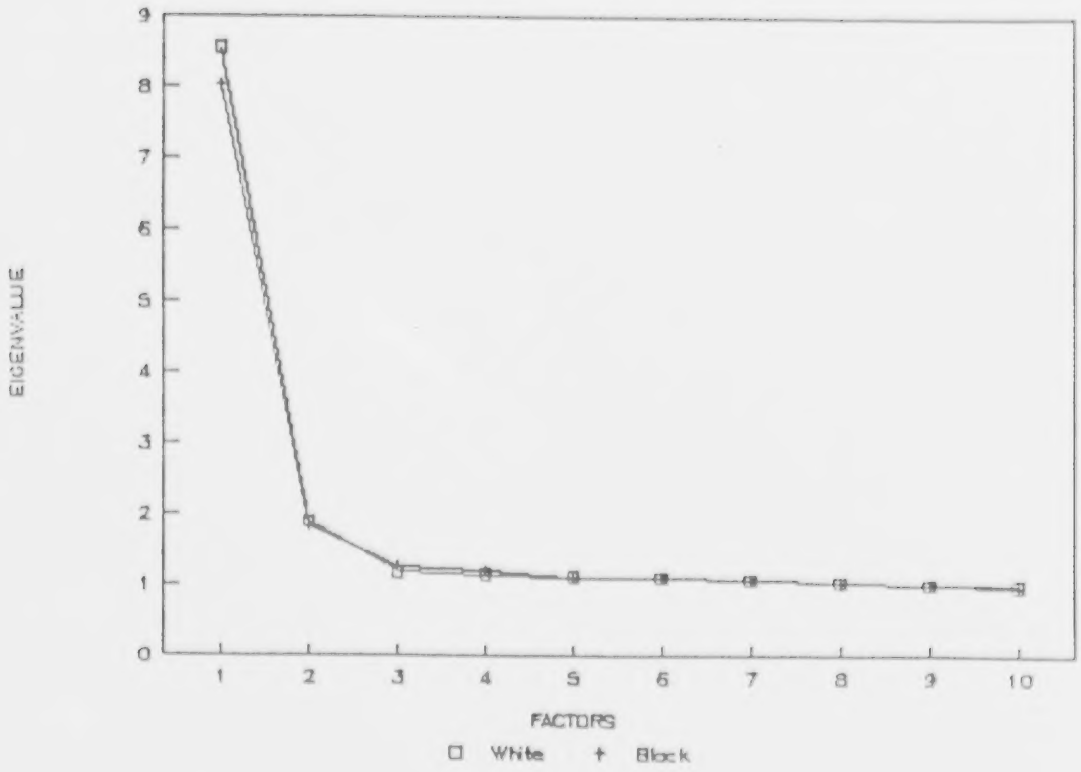

Figure 5. Scree Plot of Eigenvalues.

Table 1

Performance on Studied Item

	Correct	Incorrec	Total
Group			
Reference	Aj	Bj	nRj
Focal	Cj	Dj	nFj
Total	m1j	m0j	Tj

Table 2

Item Parameter Estimates For Selected Student Subsamples (N=1000)

	Random1			Random2			Male			Female		
Item #	a	b	c	a	b	c	a	b	c	a	b	c
1	0.498	-3.393	0.267	0.741	-2.421	0.290	0.815	-2.463	0.273	1.043	-1.771	0.215
2	0.847	0.527	0.253	0.606	0.193	0.213	0.846	0.311	0.311	0.599	0.201	0.125
3	0.958	-1.125	0.277	0.894	-1.104	0.234	0.861	-1.481	0.222	1.008	-1.444	0.201
4	0.811	0.292	0.242	0.937	0.477	0.278	0.801	0.361	0.306	0.652	-0.037	0.230
5	0.468	0.404	0.220	0.491	0.503	0.201	0.544	0.439	0.247	0.468	0.045	0.214
6	1.498	-0.306	0.349	1.410	-0.559	0.219	0.973	-0.620	0.242	1.081	-0.702	0.250
7	0.818	-2.008	0.317	0.973	-1.863	0.350	0.866	-2.219	0.282	0.923	-1.828	0.202
8	0.730	0.310	0.153	0.668	0.099	0.112	0.720	-0.126	0.112	0.716	0.101	0.113
9	1.469	0.035	0.164	1.617	0.144	0.230	1.529	-0.267	0.200	1.386	0.068	0.201
10	1.775	-0.141	0.420	1.308	-0.289	0.445	1.214	-0.580	0.211	1.027	-0.661	0.238
11	1.310	0.305	0.175	0.975	0.439	0.168	1.066	0.404	0.115	0.986	0.298	0.110
12	1.376	-0.156	0.442	1.322	-0.171	0.474	0.882	-0.759	0.264	0.819	-0.920	0.210
13	0.798	-0.341	0.236	0.715	-0.580	0.190	0.677	-0.594	0.187	0.691	-0.555	0.150
14	1.695	0.561	0.285	1.100	0.651	0.215	1.169	0.423	0.208	0.932	0.382	0.198
15	0.966	0.491	0.256	0.781	0.418	0.251	0.952	0.173	0.256	0.898	0.335	0.215
16	0.704	-1.061	0.279	0.897	-1.251	0.275	0.849	-1.381	0.210	0.928	-1.135	0.203
17	1.009	-0.608	0.292	1.190	-0.464	0.389	1.017	-0.753	0.205	1.066	-0.882	0.218
18	0.804	-0.534	0.188	0.778	-0.628	0.183	0.841	-0.577	0.222	0.917	-0.852	0.210
19	1.209	0.588	0.148	0.907	0.415	0.151	0.890	0.261	0.127	0.901	0.410	0.120
20	1.330	0.420	0.280	1.100	0.217	0.208	0.980	-0.074	0.185	0.991	0.102	0.175
21	0.606	-1.458	0.201	0.552	-1.397	0.236	0.629	-1.331	0.233	0.638	-1.389	0.215
22	1.173	-1.196	0.264	1.253	-1.283	0.181	1.128	-1.229	0.185	1.505	-1.140	0.111
23	0.589	-1.889	0.206	0.596	-2.036	0.242	0.564	-1.989	0.187	0.708	-1.909	0.106
24	0.984	-2.406	0.202	0.919	-2.756	0.228	0.930	-2.305	0.213	1.433	-2.020	0.189
25	0.830	-1.115	0.259	0.601	-1.511	0.151	0.888	-1.614	0.193	0.961	-1.252	0.168
26	0.801	-0.987	0.140	0.988	-0.700	0.198	0.837	-1.193	0.153	0.977	-0.797	0.098
27	0.780	-1.942	0.187	0.654	-2.178	0.220	0.666	-2.193	0.212	1.093	-1.808	0.201
28	1.346	-1.801	0.178	1.379	-1.701	0.199	1.129	-2.019	0.169	1.242	-1.775	0.089
29	0.999	-1.101	0.221	1.055	-1.101	0.196	0.877	-1.301	0.240	1.105	-1.153	0.201
30	0.744	-1.227	0.172	0.799	-1.011	0.229	0.890	-1.115	0.162	0.919	-1.067	0.109
31	0.955	-1.998	0.201	0.703	-2.102	0.235	0.849	-2.124	0.178	0.879	-1.796	0.125
32	0.754	-1.015	0.171	0.618	-0.913	0.239	0.764	-1.164	0.214	0.812	-1.055	0.203
33	0.993	-1.642	0.185	0.831	-1.811	0.178	1.081	-1.706	0.171	1.277	-1.448	0.108
34	0.555	-0.017	0.199	0.497	0.045	0.257	0.753	-0.052	0.250	0.678	-0.184	0.212
35	0.783	-0.762	0.188	0.874	-0.529	0.198	0.905	-0.522	0.227	0.902	-0.734	0.207
36	1.410	-0.868	0.203	1.736	-0.631	0.294	1.371	-0.998	0.160	1.482	-0.705	0.198
37	0.988	-0.318	0.173	1.292	-0.005	0.267	1.160	-0.401	0.216	1.266	-0.324	0.206
38	1.020	-0.724	0.243	1.140	-0.619	0.296	1.258	-0.790	0.254	1.154	-0.783	0.210
39	0.886	-0.716	0.241	1.093	-0.530	0.251	0.868	-0.758	0.192	0.982	-0.838	0.123
40	0.905	-0.335	0.155	1.145	-0.144	0.233	1.074	-0.282	0.180	1.021	-0.327	0.104
41	1.580	-0.049	0.201	1.348	-0.102	0.212	1.708	-0.087	0.207	1.173	-0.333	0.185
42	1.360	-0.456	0.238	1.398	-0.589	0.210	1.518	-0.736	0.208	1.472	-0.648	0.205
43	1.377	-0.189	0.095	1.148	-0.282	0.088	1.354	-0.141	0.158	1.116	-0.237	0.125
44	0.639	-0.219	0.260	0.528	-0.311	0.198	0.636	-0.233	0.243	0.597	-0.567	0.235
45	0.675	-1.315	0.201	0.583	-1.531	0.216	0.595	-1.497	0.217	0.810	-1.330	0.201

Note. a = discrimination; b = difficulty level; c = guessing.

Table 3

Item Parameter Estimates For Selected Student Subsamples (N=1000)

	Random1			Random2			White			Black		
Item #	a	b	c	a	b	c	a	b	c	a	b	c
1	0.498	-3.393	0.267	0.741	-2.421	0.290	0.735	-2.612	0.268	0.631	-2.363	0.253
2	0.847	0.527	0.253	0.606	0.193	0.213	0.757	0.491	0.287	0.635	0.598	0.243
3	0.958	-1.125	0.277	0.894	-1.104	0.234	0.788	-1.604	0.315	0.735	-1.176	0.241
4	0.811	0.292	0.242	0.937	0.477	0.278	0.994	0.181	0.275	0.960	0.660	0.307
5	0.468	0.404	0.220	0.491	0.503	0.201	0.536	0.690	0.287	0.472	0.680	0.211
6	1.444	-0.326	0.337	1.142	-0.519	0.293	1.207	-0.792	0.314	0.969	-0.048	0.359
7	0.818	-2.008	0.317	0.973	-1.863	0.350	0.935	-2.265	0.283	0.832	-1.754	0.291
8	0.730	0.310	0.153	0.668	0.099	0.112	0.804	0.203	0.245	0.855	0.540	0.179
9	1.469	0.035	0.164	1.617	0.144	0.230	1.417	-0.275	0.155	1.398	0.310	0.216
10	1.708	-0.153	0.441	1.483	-0.276	0.418	1.157	-0.687	0.304	1.235	-0.055	0.452
11	1.310	0.305	0.175	0.975	0.439	0.168	1.096	0.227	0.173	1.123	0.594	0.189
12	1.376	-0.156	0.442	1.322	-0.171	0.474	1.026	-0.492	0.382	0.789	-0.235	0.445
13	0.776	-0.393	0.223	0.724	-0.525	0.197	0.795	-0.628	0.201	1.007	-0.057	0.183
14	1.416	0.551	0.256	1.150	0.551	0.261	0.974	0.453	0.343	1.191	0.891	0.231
15	0.933	0.488	0.244	0.792	0.428	0.230	1.026	0.359	0.279	0.795	0.677	0.240
16	0.803	-1.061	0.279	0.883	-1.146	0.275	0.822	-1.713	0.234	1.129	-0.266	0.474
17	1.009	-0.608	0.292	1.190	-0.464	0.389	0.837	-1.106	0.292	0.763	-0.458	0.257
18	0.804	-0.534	0.188	0.796	-0.668	0.185	0.780	-0.930	0.218	0.658	-0.198	0.183
19	1.209	0.588	0.148	0.907	0.415	0.151	0.750	0.234	0.161	0.969	1.047	0.220
20	1.230	0.420	0.280	1.200	0.265	0.208	0.890	-0.093	0.213	1.124	0.683	0.235
21	0.606	-1.458	0.201	0.552	-1.397	0.236	0.408	-2.043	0.267	0.432	-1.297	0.258
22	1.173	-1.196	0.264	1.253	-1.283	0.181	1.214	-1.513	0.273	1.282	-0.819	0.273
23	0.589	-1.889	0.206	0.596	-2.036	0.242	0.465	-2.518	0.244	0.503	-2.115	0.231
24	0.984	-2.406	0.202	0.919	-2.756	0.228	0.640	-3.674	0.243	0.927	-2.367	0.223
25	0.830	-1.315	0.240	0.682	-1.497	0.199	0.683	-2.077	0.199	0.556	-1.157	0.221
26	0.874	-0.781	0.152	0.957	-0.701	0.168	0.786	-1.367	0.179	0.993	-0.357	0.229
27	0.780	-1.942	0.187	0.654	-2.178	0.220	0.846	-2.390	0.204	0.703	-1.863	0.211
28	1.356	-1.821	0.178	1.378	-1.811	0.196	1.018	-2.413	0.217	0.814	-2.060	0.205
29	0.999	-1.101	0.221	1.055	-1.101	0.196	0.966	-1.546	0.218	0.966	-0.934	0.213
30	0.744	-1.227	0.172	0.791	-1.084	0.209	0.704	-1.616	0.210	0.618	-0.906	0.206
31	0.955	-1.998	0.201	0.603	-2.402	0.235	0.681	-2.605	0.216	1.145	-1.650	0.208
32	0.754	-1.015	0.176	0.667	-0.934	0.238	0.924	-1.155	0.215	0.610	-0.826	0.240
33	0.993	-1.642	0.185	0.831	-1.811	0.178	0.948	-2.147	0.200	0.969	-1.555	0.236
34	0.555	-0.017	0.199	0.497	0.045	0.257	0.610	-0.452	0.160	0.628	0.191	0.213
35	0.783	-0.762	0.188	0.874	-0.529	0.198	0.972	-0.731	0.266	1.060	-0.187	0.251
36	1.410	-0.868	0.203	1.736	-0.631	0.294	1.594	-0.956	0.266	1.879	-0.490	0.334
37	0.988	-0.318	0.173	1.292	0.005	0.267	1.471	-0.127	0.325	1.307	-0.008	0.234
38	1.020	-0.724	0.243	1.109	-0.629	0.284	1.123	-0.967	0.290	1.046	-0.504	0.273
39	0.886	-0.716	0.241	1.093	-0.530	0.251	0.816	-0.993	0.203	0.918	-0.257	0.270
40	0.905	-0.335	0.155	1.145	-0.144	0.233	1.331	-0.366	0.254	1.014	0.017	0.171
41	1.580	-0.049	0.201	1.348	-0.102	0.212	1.453	-0.315	0.215	1.633	0.157	0.250
42	1.375	-0.456	0.233	1.398	-0.579	0.220	1.499	-0.735	0.281	1.515	-0.298	0.292
43	1.377	-0.189	0.095	1.148	-0.282	0.088	1.352	-0.260	0.167	1.318	0.071	0.167
44	0.619	-0.219	0.262	0.585	-0.311	0.199	0.584	-0.475	0.241	0.457	-0.067	0.202
45	0.665	-1.315	0.203	0.593	-1.473	0.213	0.665	-1.612	0.212	0.601	-1.156	0.244

Note. a = discrimination; b = difficulty level; c = guessing.

Table 4
Differences Between ICCs for Three-Parameter Model

Items	Two Random Groups		Male Vs. Female	
	Differences	Absolu.Diff	Differences	Absolu. Diff
1	0.155	0.192	0.280	0.280
2	-0.106	0.217	0.592	0.592*
3	0.098	0.098	-0.075	0.075
4	0.040	0.113	0.014	0.176
5	0.155	0.155	-0.109	0.221
6	0.157	0.245	-0.113	0.115
7	-0.007	0.043	0.187	0.187
8	-0.025	0.086	0.269	0.284
9	-0.097	0.167	0.342	0.349*
10	-0.146	0.186	-0.147	0.262
11	0.157	0.211	0.040	0.170
12	-0.090	0.101	-0.037	0.120
13	-0.049	0.089	0.122	0.122
14	0.346	0.346	0.113	0.241
15	-0.021	0.142	0.359	0.359*
16	-0.158	0.180	0.107	0.107
17	-0.144	0.144	-0.184	0.190
18	-0.054	0.063	-0.229	0.229
19	-0.135	0.233	0.280	0.280
20	0.107	0.174	0.255	0.255
21	0.004	0.104	-0.069	0.117
22	0.066	0.110	0.086	0.144
23	-0.145	0.145	-0.046	0.070
24	-0.196	0.196	-0.053	0.102
25	-0.017	0.157	0.207	0.207
26	0.087	0.124	0.381	0.381*
27	-0.130	0.167	-0.019	0.151
28	0.028	0.036	0.090	0.090
29	0.037	0.046	0.069	0.082
30	0.034	0.063	0.063	0.066
31	-0.028	0.173	0.149	0.149
32	-0.036	0.232	0.035	0.068
33	-0.092	0.139	0.138	0.138
34	-0.120	0.209	0.056	0.148
35	0.155	0.155	-0.142	0.164
36	-0.027	0.163	0.118	0.185
37	-0.011	0.206	0.105	0.105
38	-0.049	0.063	0.060	0.101
39	0.111	0.137	0.030	0.066
40	-0.053	0.134	0.191	0.191
41	-0.060	0.115	-0.116	0.269
42	-0.031	0.094	0.047	0.097
43	-0.052	0.122	0.047	0.181
44	0.129	0.129	-0.198	0.242
45	-0.132	0.173	-0.016	0.095
Mean	-0.008	0.146	0.073	0.183
S.D.	0.113	0.062	0.170	0.104
99% C.I.	(-0.30494, 0.28966)	(-0.01628, 0.30866)	(-0.37715, 0.52300)	(-0.09105, 0.457)

Table 5
Differences Between ICCs for Three-Parameter Model

Items	Two Random Groups		White Vs. Black	
	Differences	Absolu.Diff	Differences	Absolu. Diff
1	0.155	0.192	-0.078	0.113
2	-0.106	0.217	-0.097	0.165
3	0.098	0.098	0.028	0.050
4	0.040	0.113	-0.074	0.154
5	0.155	0.155	-0.046	0.111
6	0.157	0.245	0.055	0.245
7	-0.007	0.043	-0.037	0.091
8	-0.025	0.086	0.122	0.122
9	-0.097	0.167	-0.079	0.194
10	-0.146	0.186	-0.278	0.286
11	0.157	0.211	-0.113	0.120
12	-0.090	0.101	-0.290	0.346*
13	-0.049	0.089	0.081	0.131
14	0.346	0.346	0.409	0.414*
15	-0.021	0.142	0.037	0.171
16	-0.158	0.180	0.093	0.239
17	-0.144	0.144	0.156	0.156
18	-0.054	0.063	0.264	0.264
19	-0.135	0.233	0.131	0.231
20	0.107	0.174	0.195	0.211
21	0.004	0.104	0.084	0.084
22	0.066	0.110	0.071	0.071
23	-0.145	0.145	-0.112	0.112
24	-0.196	0.196	0.058	0.080
25	-0.017	0.157	0.265	0.275
26	0.087	0.124	0.231	0.232
27	-0.130	0.167	0.016	0.115
28	0.028	0.036	-0.070	0.132
29	0.037	0.046	0.026	0.053
30	0.034	0.063	0.147	0.176
31	-0.028	0.173	0.007	0.158
32	-0.036	0.232	-0.138	0.376*
33	-0.092	0.139	-0.065	0.073
34	-0.120	0.209	-0.021	0.136
35	0.155	0.155	0.047	0.047
36	-0.027	0.163	-0.201	0.201
37	-0.011	0.206	-0.035	0.237
38	-0.049	0.063	-0.026	0.077
39	0.111	0.137	0.000	0.089
40	-0.053	0.134	0.125	0.173
41	-0.060	0.115	-0.108	0.112
42	-0.031	0.094	-0.097	0.100
43	-0.052	0.122	-0.139	0.144
44	0.129	0.129	0.080	0.225
45	-0.132	0.173	-0.084	0.159
Mean	-0.008	0.146	0.012	0.166
S.D.	0.113	0.062	0.140	0.086
99% C.I.	(-0.30494, 0.28966)	(-0.01628, 0.30866)	(-0.358, 0.383)	(-0.061, 0.393)

Table 6
Summary of Mantel-Haenzel DIF Statistics across Different Sample Sizes (Male vs. Female).
(Score Group = 7)

ITEM NO	MH-CHISQ				MH-Alpha				MH-Delta			
	200	500	750	1000	200	500	750	1000	200	500	750	1000
1	0.06	1.47	1.76	1.85	0.83	1.59	1.46	1.36	0.44	-1.09	-0.89	-0.72
2	2.84	4.76	2.89	6.84*	1.49	1.37	1.23	1.32	-0.94	-0.74	-0.49	-0.65
3	1.99	0.01	0.48	2.47	0.61	0.96	0.87	0.77	1.16	0.10	0.33	0.61
4	0.18	0.00	1.43	0.77	0.89	1.02	0.86	0.91	0.27	-0.05	0.35	0.22
5	0.01	1.00	0.13	1.07	0.98	1.16	1.05	0.90	0.05	-0.35	-0.11	0.25
6	0.26	1.53	2.61	0.02	1.19	1.26	1.28	0.97	-0.41	-0.54	-0.58	0.07
7	0.86	3.12	0.44	1.75	1.58	1.83	1.19	1.33	-1.07	-1.42	-0.41	-0.67
8	0.98	1.51	3.93*	2.24	1.28	1.22	1.28	1.18	-0.58	-0.47	-0.58	-0.39
9	3.89	15.06*	17.53*	21.73*	1.71	1.92	1.77	1.78	-1.26	-1.53	-1.34	-1.36
10	0.01	1.10	0.24	0.50	1.04	0.82	0.92	1.11	-0.09	0.47	0.20	-0.25
11	0.31	0.14	4.97	0.74	0.86	0.93	0.75	0.90	0.35	0.17	0.68	0.25
12	0.38	3.13	0.00	0.45	0.82	0.73	1.00	0.91	0.47	0.74	0.00	0.22
13	0.17	0.01	0.31	0.33	0.88	0.99	1.09	1.07	0.30	0.02	-0.20	-0.16
14	0.01	0.26	0.27	0.02	0.95	1.09	1.07	0.99	0.12	-0.20	-0.16	0.02
15	0.70	1.17	4.46	13.55*	1.24	1.18	1.29	1.49	-0.51	-0.39	-0.60	-0.94
16	0.90	2.24	0.11	0.98	0.72	1.37	1.07	1.16	0.77	-0.74	-0.16	-0.35
17	0.52	0.91	0.50	0.05	1.26	0.83	0.89	1.04	-0.54	0.44	0.27	-0.09
18	5.49	2.01	12.66*	10.45*	0.52	0.78	0.60	0.67	1.54	0.58	1.20	0.94
19	4.58	0.15	1.26	0.25	1.66	0.93	1.16	1.06	-1.19	0.17	-0.35	-0.14
20	0.12	0.02	4.57	0.98	0.89	1.02	1.31	1.12	0.27	-0.05	-0.63	-0.27
21	0.01	1.73	0.80	0.87	1.06	0.78	0.86	0.88	-0.14	0.58	0.35	0.30
22	0.80	0.21	3.27	3.28	0.68	0.88	0.70	0.73	0.91	0.30	0.84	0.74
23	2.42	5.12	4.73	8.67*	0.56	0.62	0.67	0.63	1.36	1.12	0.94	1.09
24	3.25	0.44	9.72*	8.83*	0.11	0.74	0.34	0.45	5.19	0.71	2.54	1.88
25	0.01	0.01	0.87	3.05	1.07	0.98	1.19	1.32	-0.16	0.05	-0.41	-0.65
26	1.37	6.00	7.46*	3.29	1.44	1.56	1.50	1.28	-0.86	-1.05	-0.95	-0.58
27	6.17	11.29*	7.73*	6.43	0.33	0.38	0.54	0.61	2.61	2.27	1.45	1.16
28	1.75	0.06	0.13	0.31	0.48	0.88	0.89	0.86	1.72	0.30	0.27	0.35
29	2.28	1.29	0.01	0.00	0.59	0.79	0.99	1.01	1.24	0.55	0.02	-0.02
30	0.50	0.09	0.01	0.46	1.27	0.93	0.98	0.90	-0.56	0.17	0.05	0.25
31	0.01	1.00	0.01	0.22	0.98	0.72	0.99	1.11	0.05	0.77	0.02	-0.25
32	0.02	0.01	0.05	0.34	1.00	0.97	0.96	1.09	0.00	0.07	0.10	-0.20
33	0.02	0.03	0.40	0.01	1.07	1.08	1.17	0.98	-0.16	-0.18	-0.37	0.05
34	0.12	0.00	1.22	1.80	1.11	1.00	1.15	1.16	-0.25	0.00	-0.33	-0.35
35	0.01	1.22	1.82	18.61*	1.03	0.81	0.82	0.58	-0.07	0.50	0.47	1.28
36	0.02	1.98	0.70	2.60	1.03	0.73	1.13	1.29	-0.07	0.74	-0.39	-0.60
37	1.66	6.09	0.03	0.42	1.42	0.65	0.97	1.09	-0.82	1.01	0.07	-0.20
38	0.44	0.03	0.01	0.25	1.25	1.05	0.98	1.08	-0.52	-0.11	0.05	-0.13
39	0.00	7.23*	1.23	2.82	0.99	0.62	0.84	0.80	0.02	1.12	0.41	0.52
40	2.22	5.55	6.16	1.68	1.47	0.68	0.71	0.86	-0.91	0.91	0.80	0.35
41	0.16	1.51	0.86	4.36	1.14	0.80	0.97	0.77	-0.31	0.52	0.07	0.61
42	0.02	6.82*	2.43	0.84	1.00	1.71	1.31	1.15	0.00	-1.26	-0.63	-0.33
43	0.00	0.36	0.02	1.21	1.03	1.13	1.03	0.87	-0.07	-0.29	-0.07	0.33
44	0.99	0.01	2.28	5.46	0.76	1.01	0.82	0.77	0.64	-0.02	0.47	0.61
45	0.00	7.00*	0.52	1.56	0.96	0.60	0.89	0.84	0.10	1.20	0.27	0.41

* $P < 0.05$

Table 7
Summary of Mantel-Haenzel DIF Statistics across Different Sample Sizes. (White vs. Black)
(Score Group = 7)

	MH-CHISQ				MH-Alpha				MH-Delta			
ITEM NO	200	500	750	1000	200	500	750	1000	200	500	750	1000
1	1.81	0.27	1.30	0.00	2.88	0.83	1.36	0.99	-2.49	0.44	-0.72	0.02
2	0.78	7.76	2.75	5.47	0.79	0.66	0.82	0.79	0.55	0.98	0.47	0.55
3	0.03	0.00	2.78	0.63	0.90	1.03	1.34	1.13	0.25	-0.07	-0.69	-0.29
4	4.06	0.00	0.75	1.88	0.60	1.00	0.90	0.86	1.20	0.00	0.25	0.35
5	0.00	0.57	0.00	2.81	1.03	0.89	0.99	0.84	-0.07	0.27	0.02	0.41
6	0.05	0.73	3.93	2.50	0.91	1.18	1.34	1.22	0.22	-0.39	-0.69	-0.47
7	0.62	0.00	5.49	0.63	1.57	1.05	1.77	1.19	-1.06	-0.11	-1.34	-0.41
8	1.29	0.21	0.09	0.01	1.34	0.92	0.96	1.02	-0.69	0.20	0.10	-0.05
9	0.04	0.40	0.19	0.48	0.91	0.89	1.07	0.92	0.22	0.27	-0.16	0.20
10	5.39	3.54	2.46	5.31	0.50	0.70	0.79	0.75	1.63	0.84	0.55	0.68
11	0.01	0.70	12.93	4.61	1.01	0.87	0.63	0.79	-0.02	0.33	1.09	0.55
12	0.63	4.78	6.82	8.34	1.02	1.17	1.28	1.39	-0.05	-0.37	-0.58	-0.77
13	0.61	1.75	0.90	0.16	1.25	1.25	1.14	1.05	-0.52	-0.52	-0.31	-0.11
14	0.29	0.01	0.08	12.41	0.87	1.03	0.96	1.43	0.33	-0.07	0.10	-0.84
15	0.00	1.16	0.65	1.87	0.98	1.18	0.90	0.86	0.05	-0.39	0.25	0.35
16	0.85	2.24	0.29	5.93	1.39	1.36	1.11	1.42	-0.77	-0.72	-0.25	-0.82
17	1.47	0.07	1.65	4.11	1.44	1.07	1.21	1.29	-0.86	-0.16	-0.45	-0.60
18	0.01	1.31	16.55	9.54	1.01	1.21	1.70	1.42	-0.02	-0.45	-1.25	-0.82
19	3.18	0.23	0.44	1.84	0.64	1.08	0.92	1.15	1.05	-0.18	0.20	-0.33
20	0.04	5.53	4.03	4.26	1.08	1.44	1.28	1.25	-0.18	-0.86	-0.58	-0.52
21	0.01	0.26	1.79	0.99	1.02	0.90	1.23	1.14	-0.05	0.25	-0.49	-0.31
22	2.16	0.66	10.37	2.04	1.75	1.21	1.84	1.27	-1.32	-0.45	-1.43	-0.56
23	0.66	1.03	1.72	1.10	1.43	0.79	0.79	0.85	-0.84	0.55	0.55	0.38
24	0.09	0.30	0.11	0.50	1.59	0.76	0.84	1.26	-1.09	0.64	0.41	-0.54
25	9.80	4.45	8.26	16.17	3.31	1.56	1.61	1.77	-2.81	-1.05	-1.12	-1.34
26	8.61	8.42	1.84	6.60	2.44	1.71	1.23	1.39	-2.10	-1.26	-0.49	-0.77
27	0.03	7.42	1.20	2.30	1.16	2.01	1.27	1.35	-0.35	-1.64	-0.56	-0.71
28	0.35	3.67	8.00	0.84	0.68	1.95	2.29	1.25	0.91	-1.57	-1.95	-0.52
29	0.72	0.24	0.97	0.68	0.69	1.12	0.84	1.14	0.87	-0.27	0.41	-0.31
30	2.91	1.31	5.65	5.87	1.68	1.25	1.44	1.38	-1.22	-0.52	-0.86	-0.76
31	0.04	0.55	0.00	0.43	1.18	1.26	1.02	0.86	-0.39	-0.54	-0.05	0.35
32	0.20	1.12	0.83	0.03	1.06	1.22	1.26	1.16	-0.14	-0.47	-0.54	-0.35
33	0.24	0.02	0.28	0.00	1.34	1.08	0.87	1.03	-0.69	-0.18	0.33	-0.07
34	0.52	1.32	0.00	0.41	0.82	1.20	1.01	0.93	0.47	-0.43	-0.02	0.17
35	4.33	0.59	0.50	0.00	1.75	0.87	1.11	1.01	-1.32	0.33	-0.25	-0.02
36	0.16	0.49	0.36	3.27	0.85	0.85	1.12	0.77	0.38	0.38	-0.27	0.61
37	0.00	2.70	1.95	5.95	0.98	0.75	0.83	0.75	0.05	0.68	0.44	0.68
38	0.21	1.36	0.20	0.03	0.84	0.79	1.08	0.97	0.41	0.55	-0.18	0.07
39	0.25	0.03	0.03	0.06	1.19	1.05	1.03	1.04	-0.41	-0.11	-0.07	-0.09
40	0.14	0.00	0.03	0.63	0.87	0.98	0.97	1.10	0.33	0.05	0.07	-0.22
41	2.92	1.41	2.08	2.27	0.61	0.81	0.82	0.83	1.16	0.50	0.47	0.44
42	0.25	0.00	0.28	1.08	0.83	1.01	1.10	0.87	0.44	-0.02	-0.22	0.33
43	0.41	1.14	0.50	5.55	0.83	0.83	0.90	0.76	0.44	0.44	0.25	0.64
44	0.19	1.95	0.02	0.01	1.13	0.81	0.97	1.01	-0.29	0.50	0.07	-0.02
45	0.32	0.11	4.06	0.01	1.21	1.09	0.72	0.98	-0.45	-0.20	0.77	0.05

*p < 0.05

Table 8

Correlations of MH-Alpha across Different Numbers of Score Groups and Different Sample Sizes with IRT DIF Index: Male vs. Female.

IRT DIF Index		Mantel-Haenszel Alpha																								
		Score Groups=4				Score Groups=5				Score Groups=6				Score Groups=8				Score Groups=10								
Area1	Area2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
N=1000	N=1000	N=200	N=500	N=750	N=1000	N=200	N=500	N=750	N=1000	N=200	N=500	N=750	N=1000	N=200	N=500	N=750	N=1000	N=200	N=500	N=750	N=1000	N=200	N=500	N=750	N=1000	N=1000
1.000																										
0.236	1.000																									
0.087	0.434	1.000																								
-0.017	0.440	0.472	1.000																							
0.093	0.467	0.631	0.777	1.000																						
0.092	0.509	0.655	0.683	0.877	1.000																					
0.129	0.478	0.987	0.451	0.612	0.635	1.000																				
-0.014	0.451	0.464	0.995	0.773	0.689	0.441	1.000																			
0.111	0.485	0.632	0.772	0.998	0.874	0.618	0.768	1.000																		
0.099	0.506	0.646	0.689	0.884	0.999	0.626	0.695	0.881	1.000																	
0.061	0.404	0.986	0.467	0.609	0.598	0.973	0.458	0.611	0.591	1.000																
-0.022	0.414	0.389	0.974	0.723	0.645	0.360	0.986	0.715	0.651	0.382	1.000															
0.081	0.466	0.627	0.777	0.995	0.880	0.605	0.776	0.994	0.887	0.609	0.733	1.000														
0.098	0.490	0.635	0.666	0.878	0.995	0.611	0.676	0.875	0.996	0.581	0.639	0.885	1.000													
0.086	0.414	0.993	0.462	0.627	0.639	0.978	0.457	0.629	0.631	0.984	0.387	0.625	0.624	1.000												
-0.088	0.353	0.401	0.959	0.729	0.649	0.358	0.973	0.717	0.654	0.393	0.981	0.735	0.645	0.466	1.000											
0.087	0.454	0.601	0.766	0.994	0.872	0.577	0.769	0.992	0.879	0.578	0.729	0.992	0.878	0.605	0.742	1.000										
0.087	0.497	0.629	0.669	0.872	0.997	0.607	0.679	0.869	0.998	0.572	0.638	0.877	0.997	0.618	0.648	0.873	1.000									
0.090	0.453	0.996	0.463	0.632	0.663	0.986	0.459	0.635	0.654	0.981	0.387	0.632	0.648	0.994	0.398	0.607	0.641	1.000								
-0.034	0.425	0.431	0.987	0.754	0.675	0.402	0.996	0.747	0.681	0.421	0.993	0.760	0.665	0.427	0.985	0.758	0.669	0.428	1.000							
0.099	0.476	0.633	0.774	0.998	0.877	0.615	0.772	0.998	0.884	0.611	0.722	0.995	0.878	0.[illegible]	0.726	0.995	0.873	0.635	0.754	1.000						
0.089	0.497	0.631	0.675	0.876	0.998	0.608	0.684	0.872	0.998	0.574	0.642	0.881	0.997	0.619	0.650	0.875	0.999	0.642	0.673	0.877	1.000					
0.103	0.471	0.991	0.464	0.625	0.657	0.993	0.457	0.630	0.648	0.976	0.378	0.621	0.635	0.985	0.381	0.593	0.631	0.994	0.421	0.[illegible]	0.[illegible]	1.000				
-0.027	0.440	0.458	0.995	0.771	0.685	0.436	0.999	0.766	0.691	0.453	0.986	0.774	0.671	0.451	0.974	0.767	0.675	0.453	0.996	0.[illegible]	0.[illegible]	0.[illegible]	1.000			
0.101	0.491	0.636	0.771	0.997	0.871	0.623	0.768	0.998	0.877	0.618	0.714	0.993	0.871	0.632	0.717	0.991	0.866	0.637	0.747	0.[illegible]	0.[illegible]	0.[illegible]	0.[illegible]	1.000		
0.105	0.514	0.638	0.682	0.881	0.998	0.620	0.688	0.879	0.999	0.583	0.643	0.885	0.995	0.623	0.646	0.877	0.997	0.647	0.674	0.[illegible]	0.[illegible]	0.[illegible]	0.[illegible]	0.[illegible]	1.000	

Area1: Differences between ICCs for two independent random samples.
Area2: Differences between ICCs for male and female samples.

Table 9

Correlations of MH-Alpha across Different Numbers of Score Groups and Different Sample Sizes with IRT DIF Index: White vs. Black

IRT DIF Index		Mantel-Haenszel Alpha																							
		Score Groups=4				Score Groups=5				Score Groups=6				Score Groups=8				Score Groups=10				Score Groups=7			
Area1	Area2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
N=1000	N=1000	N=200	N=500	N=750	N=1000	N=200	N=500	N=750	N=1000	N=200	N=500	N=750	N=1000	N=200	N=500	N=750	N=1000	N=200	N=500	N=750	N=1000	N=200	N=500	N=750	N=1000
1.000																									
0.102	1.000																								
-0.046	-0.134	1.000																							
-0.075	0.213	0.226	1.000																						
-0.105	0.190	0.378	0.512	1.000																					
-0.014	0.173	0.478	0.630	0.577	1.000																				
-0.022	0.286	0.994	0.211	0.332	0.465	1.000																			
-0.136	0.245	0.231	0.989	0.554	0.609	0.213	1.000																		
0.123	-0.124	0.369	0.535	0.988	0.579	0.322	0.589	1.000																	
-0.036	0.116	0.515	0.652	0.591	0.986	0.499	0.642	0.610	1.000																
-0.043	0.213	0.994	0.192	0.350	0.438	0.991	0.198	0.343	0.481	1.000															
-0.123	0.212	0.237	0.990	0.551	0.623	0.223	0.991	0.578	0.650	0.285	1.000														
-0.121	-0.126	0.400	0.516	0.991	0.579	0.356	0.564	0.993	0.603	0.374	0.559	1.000													
-0.013	0.136	0.521	0.625	0.574	0.988	0.510	0.610	0.589	0.993	0.486	0.626	0.591	1.000												
-0.034	0.198	0.991	0.229	0.365	0.495	0.995	0.230	0.354	0.525	0.987	0.244	0.387	0.536	1.000											
-0.113	0.178	0.258	0.995	0.551	0.633	0.244	0.994	0.579	0.661	0.226	0.994	0.559	0.634	0.262	1.000										
-0.121	0.189	0.387	0.525	0.994	0.595	0.344	0.568	0.988	0.608	0.356	0.556	0.994	0.596	0.379	0.566	1.000									
-0.037	0.136	0.520	0.626	0.590	0.991	0.507	0.613	0.604	0.995	0.484	0.624	0.603	0.996	0.535	0.636	0.612	1.000								
-0.023	0.135	0.993	0.196	0.346	0.471	0.995	0.199	0.339	0.509	0.993	0.211	0.369	0.518	0.994	0.230	0.356	0.516	1.000							
-0.108	0.134	0.064	0.965	0.498	0.531	0.040	0.967	0.520	0.569	0.044	0.850	0.495	0.540	0.058	0.934	0.502	0.539	0.044	1.000						
-0.121	0.213	0.362	0.533	0.994	0.573	0.317	0.582	0.996	0.596	0.334	0.576	0.995	0.578	0.350	0.575	0.995	0.593	0.332	0.518	1.000					
-0.042	0.123	0.520	0.634	0.586	0.981	0.505	0.627	0.610	0.997	0.487	0.635	0.604	0.994	0.531	0.646	0.606	0.996	0.516	0.556	0.594	1.000				
-0.027	-0.132	0.992	0.203	0.347	0.479	0.995	0.205	0.340	0.515	0.991	0.217	0.370	0.524	0.995	0.237	0.359	0.523	0.999	0.048	0.334	0.522	1.000			
-0.138	0.134	0.280	0.985	0.567	0.620	0.262	0.992	0.601	0.659	0.251	0.991	0.579	0.629	0.279	0.995	0.581	0.632	0.251	0.960	0.595	0.648	0.257	1.000		
-0.102	-0.121	0.380	0.527	0.992	0.578	0.334	0.578	0.999	0.607	0.355	0.570	0.996	0.588	0.365	0.570	0.992	0.601	0.360	0.515	0.998	0.607	0.351	0.512	1.000	
-0.044	0.136	0.523	0.641	0.594	0.987	0.509	0.632	0.612	0.997	0.489	0.642	0.609	0.995	0.535	0.652	0.615	0.997	0.518	0.562	0.600	0.998	0.524	0.651	0.610	1.000

Area1: Differences between ICCs for two independent random samples.
Area2: Differences between ICCs for male and female samples.

Table 10
Agreements in Detecting DIF on the MFTC Between MH and IRT Results.
(Male vs. Female)

IRT Results:	Biased		Unbiased		Agreement	
MH Results:	B	NB	B	NB	P1	P2
MH (n=200)	----	4	----	41	91%	----
MH (n=500)	1	4	3	37	84%	2%
MH (n=750)	2	2	3	38	89%	4%
MH (n=1000)	3	1	4	37	89%	7%

Note: For IRT Method a Sample Size of 1000 was used.
P1: Percent of test items identified consistently as "biased" or "unbiased" in two DIF methods.
P2: Percent of test items identified as "biased" in two DIF methods.

Table 11
Agreements in Detecting DIF on the MFTC Between MH and IRT Results.
(White vs. Black)

IRT Results:	Biased		Unbiased		Agreement	
MH Results:	B	NB	B	NB	P1	P2
MH (n=200)	----	3	2	40	89%	----
MH (n=500)	1	3	2	40	91%	----
MH (n=750)	1	2	5	37	89%	2%
MH (n=1000)	2	1	2	40	93%	4%

Note: For IRT Method a Sample Size of 1000 was used.
P1: Percent of test items identified consistently as "biased" or "unbiased" in two DIF methods.
P2: Percent of test items identified as "biased" in two DIF methods.

References

Angoff, W.H. (1972). A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R.A. Berk (Ed.), Handbook of methods for detecting test bias. Baltimore, MD: Johns Hopkins University Press.

Angoff, W.H., & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-105.

Baghi, H. (1988). The stability of item parameter estimates and item characteristic curves across time using different item response models and different estimation procedures. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Baghi, H., & Ferrara, S. (1989, March). A comparison of IRT, Delta Plot, and Mantel-Haenszel techniques for detecting differential item functioning across subpopulations on the Maryland Test of Citizenship Skills. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Burrill, L.E. (1982). Comparative studies of item bias methods. In R.A. Berk (Ed.), Handbook of Methods for Detecting Test Bias. Baltimore: Johns Hopkins University Press.

Cleary, T.A., & Hilton, T.L. (1968). An investigation of item bias. Educational and Psychological Measurement, 28, 61-75.

Cox, D.R. (1970). Analysis of binary data. London: Methuen and Co., Ltd.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. NY: Holt, Rinehart and Winston.

Hills, J.R. (1981). Test item bias detection methods: 1977 to 1980. Unpublished monograph, Florida State University.

Hills, J. R. (1989). Screening for potentially biased items in testing programs. Educational Measurement: Issues and Practice, 8, 5-11.

Hills, J.R., and King, F.J. (1982). Detecting biased items in minimal competency tests. Paper presented at the Fifth International Symposium on Educational Testing, Stirling, Scotland.

Hills, J.R., and Maruhnich, N. (1977). Test item bias analysis: An analysis of methodology and recommendations. Unpublished monograph, Florida State University.

Hambleton, R.K., and Murray, L. (1983). Some goodness of fit investigation for item response models. In R.K. Hambleton (Ed.), Applications of item response theory. Vancouver: Educational Research Institute of British Columbia.

Harris, D.J., and Hoover, H.D. (1986). The stability of selected item bias indices. Paper presented at the National Council on Measurement in Education, San Francisco.

Harris, D.J., and Kolen, M.J. (1986). A comparison of the stability of item bias indices using the bootstrap. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Holland, P.W., & Thayer, D.T. (1986). Differential item performance and the Mantel-Haenszel procedure. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Holland, P. W., & Thayer, D.T (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), Handbook of Methods for Detecting Test Bias. Baltimore: Johns Hopkins University Press.

Lord, F.M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Poortinga (Ed.), Bias problems in cross-cultural psychology. Amsterdam: Swets and Zeitlinger.

Merz, W., & Grossen, N. (1978). An empirical investigation of six methods for examining test item bias. Washington, DC: National Institute of Education.

Mislevy, R.J., & Bock, R.D. (1986). PC-Bilog. [Computer program.] Moorseville, IN: Scientific Software.

Osterlind, S.L. (1987). Test item bias: Quantitative applications in social sciences. Beverly Hills, CA: Sage Publications.

Phillips S.E., and Mehrens, W.A. (1988). Comparison of methods for detecting differential item performance due to instructional test misalignment. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Raju, N.S., Bode, R.K., & Larsen, V.S. (1989). Am empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement on Education, 2 (1), 1-13.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Rudner, L.M. (1977). An evaluation of select approaches for biased item modification. Unpublished doctoral dissertation, Catholic University of America, Washington, DC.

Rudner, L.M., Getson, P.R., & Knight, D.L. (1980). Biased item detection techniques. Journal of Educational Statistics, 5, 213- 233.

Scheuneman, J.D. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Shepard, L.A., Camilli, G., & Averill, M. (1981). Comparisons of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, L.A., Camilli G., and Williams, D. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22 (4), 77-105.

Skaggs, G., & Lissitz, R. (1988). The effect of estimation methods on the stability of selected item bias indices. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Subkoviak, M.J., Mack, J.S., Ironson, G.H., & Craig, R.D. (1984). Empirical comparison of selected item detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

Wright, B.D., Mead, R., & Bell, S. (1979). BICAL: Calibrating items with the Rasch model. (Report No. RM23b). Chicago: Statistical Laboratory, Department of Education, University of Chicago.

dif90